Big data and Stratified medicine — What does it mean for children?

Ding WY[1], Beresford MW[2,3], Saleem MA*[1,4], Ramanan AV*[5,6]

Affiliations:

1. Bristol Renal, Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
2. Institute of Translational Medicine, University of Liverpool, Liverpool, UK
3. Department of Paediatric Rheumatology, Alder Hey Children's NHS Foundation Trust, Liverpool, UK
4. Department of Paediatric Nephrology, Bristol Royal Hospital for Children, University Hospitals Bristol NHS Foundation Trust, Bristol, UK
5. Department of Paediatric Rheumatology, Bristol Royal Hospital for Children, University Hospitals Bristol NHS Foundation Trust, Bristol, UK
6. Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

*These authors contributed equally

Corresponding author:
Professor A. V. Ramanan FRCPCH, FRCP
Consultant Paediatric Rheumatologist
Bristol Royal Hospital for Children & Royal National Hospital for Rheumatic Diseases, Bath
Upper Maudlin Street, Bristol, BS2 8BJ
Tel: +44 (0) 117 342 0149
Email: avramanan@hotmail.com

Word Count: 3,409 words

ABSTRACT

Stratified medicine in paediatrics is increasingly becoming a reality, as our understanding of disease pathogenesis improves and novel treatment targets emerge. We have already seen some success in paediatrics in targeted therapies such as cystic fibrosis for specific CFTR (cystic fibrosis transmembrane conductance regulator) variants. With the increased speed and decreased cost of processing and analysing data from rare disease registries, we are increasingly able to use a systems biology approach (including '-omics') to screen across populations for molecules and genes of interest. Improving our understanding of the molecular mechanisms underlying disease, and how to classify patients according to these will lead the way for targeted therapies for individual patients. This review article will summarise how 'big data' and the 'omics' are being used and developed, and taking examples from paediatric renal medicine and rheumatology, demonstrate progress being made towards stratified medicine for children.

INTRODUCTION

Stratified medicine aims to define distinct patient subgroups based on increased understanding of the pathophysiology of disease, and ultimately, will enable tailoring of management and therapies to each individual patient. Precision medicine and personalised medicine are sometimes used interchangeably with stratified medicine but there are some subtle differences. Precision medicine, in addition to the above, encompasses the repeated monitoring of patients to enable fine-tuning of treatment to patient response.[1] Personalised medicine, in addition to biological stratification, takes into account patient choice and participation.[2] Currently we already clinically stratify patients, usually from observational criteria, in order to manage their conditions, but we are now becoming armed with a wealth of additional data that could provide us with more clarity and precision via a systems biology approach.[3]

For certain diseases such as paediatric rare renal diseases and juvenile idiopathic arthritis, clinical research patient registries (e.g. RaDaR, the National Registry of Rare Kidney Diseases[4], British Society for Paediatric and Adolescent Rheumatology Biologics Registries[5,6] and Childhood Arthritis Prospective Study (CAPS)[7]) are well established in the UK. These have provided a bank of phenotypical data and biological samples for study of these disorders. With the increasing affordability and speed of processing these samples, and vast improvements in our ability to analyse this data, we are now well placed to further stratify our paediatric patients for targeted therapy. NHS England is aiming for improved diagnostics via personalised medicine by 2020, and for improved targeted therapies by 2025.[2] The 100,000 genomes project has been launched with the view of achieving these aims for rare disease and cancer.[8] Here, we aim to discuss broader progress in stratified medicine and how it will be particularly relevant to rare diseases in paediatrics.

RARE DISEASES IN CHILDREN

Rare disease in Europe is defined as having a prevalence of less than 50 in 100,000 cases.[9] There are between 5000 to 8000 rare diseases[10], which

affects between six to eight per cent of the population in total. Approximately 75% of rare diseases affect children, and 30% of these children die by the age of five.[11] 80% of rare disease is thought to be genetic in origin, while the remaining conditions have varying aetiologies, for example, infection, toxins, autoimmunity, and certain cancers.

Classification of rare disease currently is often observational, and does not necessarily reflect underlying biological mechanisms. For example, in nephrotic syndrome, current classification of disease is based on histology and patient responsiveness to steroids. This has limited utility in prognostication and in deciding appropriate therapeutic interventions for patients. With advances in systems biology approaches (the computational and mathematical modelling of complex biological systems, using a holistic approach), the identification of biomarkers, or important differentiating phenotypical features, will help to stratify patients in a more meaningful manner.

Patient registries

In order to better understand each specific rare disorder better, many groups of clinicians and investigators interested in these rare diseases have come together to establish registries. In the UK, some examples are NephroS (The nephrotic syndrome study)[12], NURTuRE (National Unified Renal Translational Research Enterprise)[13], UK Juvenile-onset SLE Cohort Study and Repository[14], BCRD (Biologics for Children with Rheumatic Diseases)[5], CAPS (Childhood arthritis prospective study)[7] and EPIPEG (Epilepsy in infancy: relating phenotype to genotype)[15]. Due to the rarity of each condition, it is essential that every patient possible be recruited to ensure that we have a good representation of biological samples to analyse. For each study, detailed clinical phenotypical data as well as a range of biological samples are collected at various time points in the course of the disease. Often these registries have data collection integrated into routine clinical care. These produce rich deep clinical and biological phenotypic data that will be analysed to generate molecular and genetic diagnostic criteria that will reflect underlying disease mechanisms. (Figure 1) It is important that in future clinical trials (both

commercially sponsored and investigator-initiated studies) that associated biological samples are collected to enable studies seeking to identify biomarkers for disease response or biomarkers associated with development of side effects.

A SYSTEMS BIOLOGY APPROACH

Traditionally, science has had a reductionist approach in identifying specific perturbations in biological pathways in various disease processes. Although this has had some success, particularly for single gene disorders, the regulation of biological processes is complex and manifold, and it stands to reason that a more holistic systems biology approach would be more successful in identifying molecular and genetic diagnostic criteria. Disease onset often has specific predisposition factors and/or triggers, be they genetic or environmental, and each trigger might result in a cascade of events that affect more than one biological network.

The ability to generate vast amounts of data from the ‘omics’ technologies — genomics, transcriptomics, proteomics, and metabolomics—will enable identification of perturbations in biological networks on a systems biology level. This has been helped by the increased speed and decreased cost of processing these biological samples, and the advancement in bioinformatics so that we can better analyse these samples. Simplistically, differences or patterns can be compared between two or more clinical states, e.g. recurrence of disease versus no recurrence, or response to a therapy, and these biological samples are analysed for differences in molecules (e.g. RNA for transcriptomics).

Hood et al[3] propose viewing biology as an information science, which is necessary when processing the extensive datasets that traditional data architectures are unable to efficiently handle, also known as ‘big data’ (see Glossary)[16] that is generated by the ‘omics’. Techniques to analyse this big data often involves using ‘machine learning’ algorithms, using either supervised or unsupervised learning (see below). Molecules of interest can be checked for statistical and biological relevance. They can be grouped, for example by the

function of these molecules, to check whether there are patterns in ‘networks’ (based on previous data available—these can be checked against databases e.g. STRING[17]) for a specific disease state. These networks might then point to a specific biological mechanism underpinning disease, which would then give us clues for the pathophysiology of disease. (Figure 2) Also, we are able to test specific molecules within the pathways that might be markedly changed in diseased states, and from these, we can determine molecules that might have the best specificity and sensitivity for predicting relapses or disease activity, for example. Using more traditional laboratory methods, the role of these newly identified molecules or biomarkers can then be confirmed.

Genomics

In order to study the genome, most studies were dependent on either microarrays or Sanger sequencing. Microarrays require prior knowledge of genetic sequences, while Sanger sequencing is expensive and relatively slow. Over the last decade, genomics has progressed tremendously due to the introduction of Next Generation Sequencing (NGS) technologies.[18]

In traditional Sanger sequencing, DNA is denatured, then annealed to a primer of choice, and elongated by the addition of dNTPS (deoxynucleotides—cytosine, guanine, adenosine and thymine). The substitution of dideoxynucleotides (ddNTPs) for deoxynucleotides (dNTPs) for one of these bases, e.g. cytosine, results in early termination of the DNA strand at that particular ddNTP. These shortened DNA strands can be detected by capillary electrophoresis and the position of the base on the strand determined according to the length of the strand. This process is laborious and the first big improvement was the introduction of fluorescent tags to the ddNTPs, which enabled automated detection of different fluorescent tags. NGS (of which there are several types) has taken the steps from Sanger sequencing—sequencing, electrophoresis and detection—and combined it into an array-based system where the reactions are fixed on a solid support on a plate.[18] Millions of sequencing reactions can then be processed in parallel, and detection is automated. As a result, NGS has resulted in a much higher throughput with

increased speed and decreased costs, but with a corresponding increase in its error rate.[19] As such, Sanger sequencing is used as an established and validated technique, to independently check specific sequence changes discovered by NGS.

NGS for whole genome sequencing currently costs approximately USD1000 per person[20], with Illumina's CEO (one of the largest companies producing NGS technologies) predicting that this cost will reduce to USD100 within the next 10 years. This will be particularly relevant in the context of diagnosis in rare disease. Although the cost and speed of sequencing is rapidly decreasing, there are further limitations in the costs of storing and handling this information, as well as the analysis and interpretation of such information.

NGS has enabled many research studies to carrying out whole exome sequencing (WES) or whole genome sequencing (WGS) (see Glossary). We have seen much success in identifying genes for rare disease in NephroS, the nephrotic syndrome study. Using NGS technology for whole exome sequencing (WES) of patients with steroid-resistant nephrotic syndrome, we have identified several new genes causative for steroid-resistant nephrotic syndrome including *MAGI2* and *FAT1*[21,22]. This has contributed to the development of the gene panel offered by Bristol Genetic Laboratories, which includes 70 genes for nephrotic syndrome in a turnaround time of 6 weeks.[23] This has already changed clinical practice, with gene panel testing becoming standard for any newly diagnosed patient, which potentially avoids the need for renal biopsy, as well as avoiding non-specific immunosuppression regimes.[24] In addition, combining genomic and clinical profiling of this cohort of children with steroid-resistant nephrotic syndrome has enabled stratification of patients according to aetiology, and hence facilitated appropriate management for individual patients e.g. according to high or low risk for recurrence of disease post-renal transplantation.[25]

Pharmacogenomics is a field where genetic sequence is used to inform drug development, selection and dosing, as well as predict side effects.[26] For example, a recent genome wide association study (GWAS) (see Glossary) in

children with asthma discovered a genetic variation in the platelet-derived-growth-factor-D (PDGFD) locus that increased the risk of adrenal suppression in response to steroid therapy.[27] This finding will potentially enable risk stratification of patients for more intensive monitoring of adrenal function, as well as encourage use of alternative medications e.g. leukotriene antagonists or anti-IgE therapy in these patients where possible.[27]

Cystic fibrosis has seen great success from targeted treatments for specific cystic fibrosis variants (see Glossary). Ivacaftor is a CFTR potentiator that increases the probability of the channel opening in CFTR gating mutations.[28] Lumacaftor is a CFTR corrector that corrects CFTR misprocessing and mislocalisation in the p.Phe508del mutation, and in combination with Ivacaftor has been shown to reduce pulmonary exacerbations in patients with this variant.[29] This is the end result of identifying the correct pathogenic variants, understanding the effect of the mutations of the CFTR protein, and using targeted therapies to correct this. With increasing understanding of not only genomics, but also protein function and localization, other rare diseases can learn from the cystic fibrosis story in targeting treatments in genetic diseases.

Transcriptomics

Transcriptomics is the study of the entire transcriptome i.e. RNA. This is valuable in not only giving us information about coding RNA i.e. RNA which will be translated into proteins, but also about noncoding RNA e.g. small interfering RNA, microRNA, which have functions in regulating gene expression. Transcriptomics currently is carried out via 2 methods—microarray or RNA sequencing (RNA-seq). For microarrays, oligomeric probes are fixed on a plate, which represent the whole or majority of the known transcriptome. Fluorescently labelled transcripts (from the sample of interest) are applied to the plate, bind to their complementary probe, and the fluorescent output measured. This requires prior knowledge of sequences in order to generate appropriate probes, and its dynamic range is limited by the minimal fluorescence detectable and fluorescence saturation. RNA-seq involves

reverse transcription of RNA to complementary DNA, and next generation sequencing of complementary DNA. RNA-seq has a higher sensitivity and dynamic range, but data analysis is also more complex and labour intensive.[30] RNA-seq also requires very little RNA, and can be carried out on single cells.[31] This might be particularly relevant for cancer where there is a heterogeneous population of cells within a tumour, and is being further developed in spatial transcriptomics to visualize transcriptome variation within a tissue biopsy sample.[32]

One example where transcriptomics has been highly informative has been in lymphocyte transcriptomics in autoimmunity (e.g. SLE, ANCA-associated vasculitis) and chronic infections in adults.[33] This has revealed that a pattern of CD8 T-cell exhaustion is predictive of reduced relapses in autoimmunity, but is associated with poor clearance of chronic infection. This is inversely correlated with CD4 co-stimulation. From the data, they identified one marker, KAT2B, which could potentially have utility as a surrogate marker of clinical outcome in these diseases.[33] This will enable prognostication of clinical course, and also has identified potential therapeutic targets to reduce relapse in autoimmune disease. Similarly, a landmark study in paediatric SLE demonstrated that signature patterns could be derived from lymphocyte transcriptomics from children, using longitudinal profiling.[34] This ‘personalised immunomonitoring’ revealed seven distinct patient groups according to disease activity, supported by genotypes. This has clear implications for tailored therapies, and trial design.

Proteomics

Proteins essentially control all cellular functions. Proteomics enables the study of the entire proteome, the entire protein constituents of cells. Proteins are enzymatically broken down e.g. by trypsin into peptides, and these peptides are separated by liquid chromatography and detected by mass spectrometry according to their size and charge. The abundance of each peptide can be measured and these can be mapped to its original protein.

Proteins are complex and are subject to many post-translational modifications that determine its three-dimensional structure, cellular localization, biological function or whether it is targeted for degradation. Some examples of these post-translational modifications include phosphorylation, glycosylation, acetylation, and ubiquitylation.[35] These modifications lead to a change in mass, which can be detected by mass spectrometry. Phosphoproteomics, the study of phosphorylation of the proteome, can be valuable in identifying phosphorylation events important in the pathogenesis of disease.

Of particular interest is investigation of the plasma proteome for biomarkers. However, this has proven challenging due to the complexity of the plasma proteome and its inherent variability across the population.[35] Plasma contains many high and medium abundance proteins, while biomarkers are more likely to be lower abundance proteins. This requires extensive processing of the sample e.g. reducing high abundance proteins by immunodepletion and fractionation of samples, which increases variability and decreases throughput.[36] However, improved techniques now enable detection of greater number of proteins, while requiring a smaller starting plasma volume.[36,37] The myeloid related protein (MRP) complex 8/14 (S100A8/9, also known as calprotectin) has been shown to be marker of response to treatment in children with juvenile idiopathic arthritis and also a marker of flare after cessation of treatment.[38–40] Improvements in proteomic technology will aid in the search for further clinically relevant biomarkers, particularly from easily accessible fluids such as plasma.

Using targeted proteomics, a breakthrough study identified the M-type phospholipase A_2 receptor (PLA_2R) as the antigenic target in a large majority of idiopathic membranous nephropathy.[41] This established membranous nephropathy as an autoimmune disease, where antibodies against PLA_2R are found in approximately 70% of patients with idiopathic membranous nephropathy. This potentially circumvents the need for kidney biopsy in high risk patients who are positive for PLA_2R antibodies, and has potential utility in predicting treatment outcomes and guiding treatment.[42]

Metabolomics

Metabolomics is the study of a range of metabolites and low molecular weight molecules, and is very much a field still in its infancy. The metabolome consists of diverse molecules including peptides, lipids, amino acids, nucleic acids, carbohydrates, organic acids, vitamins, minerals, food additives, drugs, toxins, pollutants and any chemical with a molecular weight less than 2,000 daltons.[43] The HMDB (Human Metabolome Database) contains in excess of 40,000 annotated metabolite entries.[43] Detection methods largely use chromatography and mass spectrometry, or nuclear magnetic resonance spectroscopy; but due to the large variety of molecules in the metabolome, these protocols are not standardised and vary from laboratory to laboratory.[44] Clinically, we currently use small molecules to diagnose inborn errors of metabolism, for example, phenylalanine for phenylketonuria. Our metabolome is partially dependent on environmental exposure e.g. via gut microbiota, and not solely dependent on our genome. Some of these small molecules might change in response to environmental triggers, and metabolomics could have utility in understanding pathogenesis of disease that have significant environmental triggers.

Early work using metabolomics show that urinary metabolomics at birth could be a useful tool to identify premature infants who are at high risk of developing chronic lung disease.[45] In addition, a metabolomics approach has also identified a metabolic signature early in pregnancy that could potentially predict for small for gestational age infants.[46] These studies have identified some biologically relevant molecules that might be useful biomarkers, but will require validating in larger cohorts.

Data Analysis – developing new tools

Bioinformatics is essential for management of data in modern biology and medicine. Bioinformatics is defined as the application of tools of computation and analysis to the capture and interpretation of biological data (ref Bayat, BMJ. 2002 Apr 27; 324(7344): 1018–1022.). It is an interdisciplinary field, which harnesses computer science, mathematics, physics, and biology.

The challenge of handling and organising huge amounts of data, and interpreting them to identify signature patterns that correspond to disease mechanism and behaviour is a considerable one. To complement existing bioinformatics approaches, machine learning tools are increasingly being developed and utilised, adapting innovations from the world of artificial intelligence and computing/mathematics.

There are now a number of state-of-the-art predictors for estimating the functional impact of genetic variation in human disease (e.g. Ensembl Variant Effect Predictor[47]). Based on data integration algorithms from machine learning, tools have been developed for predicting the functional impact of single nucleotide variants (SNVs), indels (short insertions and deletions of genetic code) and haplo-insufficiency, as examples. These tools use sequence and other types of data, such as data drawn from ENCODE (Encyclopedia of DNA elements)[48].

Further innovation is being made to integrate broader datasets (e.g. transcriptomics, epigenetics). This includes so-called unsupervised learning based on Bayesian methods (these maximise the probability of the model given the data). The approach is called Latent Process Decomposition (LPD)[49], and assesses the structure of a dataset in the absence of knowledge of clinical outcome or biological role (hence unsupervised), and has the benefit of being more objective. These are mixed membership models, that is, a data sample (e.g. derived from a patient) is represented as a combinatorial mixture over underlying functional states. This concept is very important in many biomedical contexts and contrasts with the hierarchical cluster analysis (dendrograms) commonly used by biologists in which data samples or genes are uniquely assigned to clusters. Hierarchical clustering restricts the number of dimensions that can be taken into account and restricts genes to one cluster, while using LPD enables many dimensions to be taken into account and has no such restriction on genes. Thus, in the context of cancer, for example, there is frequent heterogeneity between cell types present; LPD is able to effectively

model for underlying causative mechanisms. In recent work, LPD has been used to isolate the signature of aggressive prostate cancer against nonaggressive disease[50], showing its utility in predicting individual disease outcomes.

FUTURE OUTLOOK

We have seen some success in using big data to develop stratified medicine in paediatrics, where improved diagnostics has helped to understand pathogenesis of disease and stratify patients for targeted therapies. With improved understanding of the molecular networks underpinning disease, we will be able to identify the best targets for therapy and have a better understanding of potential effects of these therapies at a systems level. This will enable us to provide individualised therapy for enhanced efficiency and safety.

As next generation sequencing technologies progress, we are already seeing this coming into clinical use as it becomes more affordable and rapid. For example, whole genome sequencing within 24 hours for sick neonates might become an important diagnostic tool in diagnosing genetic disorders. With the identification of further biomarkers enabling stratification, we will be able to target specific patients for clinical trials of new therapies, and reduce the number of patients exposed to the ‘wrong’ medications. For rare disease, often patient numbers are small, but we can build on previous meta-analyses and the information from big data to design trials that are powered appropriately, using a Bayesian approach.

This is further becoming a reality in the UK, with two major new paediatric projects underway in Stratified Medicine based on existing and ongoing collection of patients to specific cohorts, the NURTuRE cohort in renal medicine[13], and the CLUSTER Consortium: childhood arthritis and its associated uveitis; stratification through endotypes and mechanism to deliver benefit[51]. These aim to build and embed new methodologies, alongside comprehensive systems biology analyses of clinically deeply phenotyped and

longitudinally followed patient cohorts. The ultimate aim is re-definition of specific rare diseases in these areas according to molecular signatures.

As we become enabled by an increase in our depth of the understanding of the molecular mechanisms of disease, we will be increasingly empowered to stratify our patients such that the right person receives the right treatment at the right time.

Figure 1. Stratified medicine approach. Patient registries will enable collection of phenotypical data and biological samples. These will be processed and analysed using supervised and unsupervised approaches as appropriate. This will enable stratification of patients into subgroup according to their molecular signature, into more meaningful and biologically relevant subgroups, paving the way for clinical trials targeting the correct subgroup. Patients will receive more targeted therapies and personalised prognostic information.

Figure 2. A simple overview of the typical workflow involved in sample processing using '-omics'. Patient samples are processed using 'omics' technology. Data are often presented using heat maps after hierarchical clustering analysis. Heat maps (above adapted from Stubbs et al 2018[52] with kind permission of author) give an overall pictorial representation of upregulation or downregulation (red being up regulation and blue being down regulation in this particular heat map) of various genes, for example. Functional protein networks are used to analyse and understand the significance of the data. Novel and relevant molecules of interest are then confirmed by traditional laboratory methods. The data will inform and enable novel classifications of disease.

Glossary	
Big data	Extensive datasets that traditional data architectures are unable to efficiently handle. Characteristics of these datasets include increased volume, variety, velocity and variability.[16]
Genetic variants	Used to refer to a specific region of the genome that differs between 2 genomes. Some variants result in phenotypically significant effects that result in illness while some are not disease causing. Genetic variants are reported clinically as 'benign', 'likely benign', 'uncertain significance', 'likely pathogenic' and 'pathogenic'.[53]
Genome wide association study (GWAS)	Study of genetic variants (typically SNPs—single nucleotide polymorphisms) and their associations with disease. Usually, the SNPs are identified using SNP microarrays.
Mutation	Permanent alteration in DNA sequence.
Whole exome sequencing (WES)	Sequencing of the entire exome i.e. the protein coding area of DNA that comprises about 1% of the total genome.
Whole genome sequencing (WGS)	Sequencing of the entire genome, including intronic regions.

Contributors: WYD wrote the first draft, revised the manuscript and approved the final version. AVR and MAS conceived the work, edited the manuscript and approved the final version. MWB edited the manuscript and approved the final version.

Funding: WYD is funded by a Kidney Research UK Clinical Research Fellowship.

Competing interests: MAS has external consultancy roles with UCB, Retrophin and Pfizer.
AVR has received Speaker fees/Consultancy for Abbvie, Eli Lilly, UCB and SOBI.

REFERENCES

1 Day S, Coombes RC, McGrath-Lone L, *et al.* Stratified, precision or personalised medicine? Cancer services in the 'real world' of a London hospital. *Sociol Health Illn* 2017;**39**:143–58. doi:10.1111/1467-9566.12457

2 Graham, Ellen (Medicines, Diagnostics and Personalised Medicine Unit, Medical Directorate NE. IMPROVING OUTCOMES THROUGH PERSONALISED MEDICINE Working at the cutting edge of science to improve patients' lives. 2016.

3 Hood L, Balling R, Auffray C. Revolutionizing medicine in the 21st century through systems approaches. *Biotechnol J* 2012;**7**:992–1001. doi:10.1002/biot.201100306

4 Registry (RaDaR) | Rare Renal. http://rarerenal.org/radar-registry/ (accessed 27 Mar 2018).

5 Biologics for Children with Rheumatic Diseases. http://www.bcrdstudy.org/default.asp (accessed 17 May 2018).

6 British Society for Paediatric and Adolescent Rheumatology Etanercept

Registry. https://www.rheumatology.org.uk/Knowledge/Registers/Juvenile-Idiopathic-Arthritis-register (accessed 17 May 2018).
7 Childhood Arthritis Prospective Study (CAPS). https://www.caps-childhoodarthritisprospectivestudy.co.uk/ (accessed 17 May 2018).
8 The 100,000 Genomes Project | Genomics England. https://www.genomicsengland.co.uk/the-100000-genomes-project/ (accessed 4 Apr 2018).
9 Field M, Boat T, editors. Profile of Rare Diseases. In: *Rare Diseases and Orphan Products: Accelerating Research and Development.* Washington (DC): : National Academies Press (US) 2010. 41–72.
10 European Medicines Agency - Overview - Orphan designation. http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000029.jsp&mid=WC0b01ac0580b18a41 (accessed 27 Mar 2018).
11 What is a rare disease? | Great Ormond Street Hospital Children's Charity. https://www.gosh.org/what-we-do/research/zayed-centre-research-rare-disease-children/rare-diseases/what-rare-disease (accessed 27 Mar 2018).
12 NephroS Study Information | Rare Renal. http://rarerenal.org/clinician-information/nephrotic-syndrome-clinician-information/nephros-study/ (accessed 20 May 2018).
13 NURTuRE – A unique kidney biobank. https://www.nurturebiobank.org/ (accessed 17 May 2018).
14 UK JSLE Study Group - Institute of Translational Medicine - University of Liverpool. https://www.liverpool.ac.uk/translational-medicine/research/ukjsle/ (accessed 17 May 2018).
15 EPIPEG – Epilepsy in infancy: relating phenotype to genotype. http://epipeg.co.uk/ (accessed 17 May 2018).
16 NIST Big Data Public Working Group. NIST Special Publication 1500-1 - NIST Big Data Interoperability Framework: Volume 1, Definitions. *NIST Spec Publ* 2015;**1**:32. doi:http://dx.doi.org/10.6028/NIST.SP.1500-1
17 STRING: functional protein association networks. https://string-db.org

18 Metzker ML. Sequencing technologies the next generation. *Nat Rev Genet* 2010;**11**:31–46. doi:10.1038/nrg2626

19 Schirmer M, Ijaz UZ, D'Amore R, *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* 2015;**43**. doi:10.1093/nar/gku1341

20 DNA Sequencing Costs: Data - National Human Genome Research Institute (NHGRI). https://www.genome.gov/sequencingcostsdata/ (accessed 21 May 2018).

21 Bierzynska A, Soderquest K, Dean P, *et al.* MAGI2 Mutations Cause Congenital Nephrotic Syndrome. *J Am Soc Nephrol* 2017;**28**:1614–21. doi:10.1681/ASN.2016040387

22 Gee HY, Sadowski CE, Aggarwal PK, *et al.* FAT1 mutations cause a glomerulotubular nephropathy. *Nat Commun* 2016;**7**:10822. doi:10.1038/ncomms10822

23 Roberts E, Williams M, Watson E. Renal panel for Steroid Resistant Nephrotic Syndrome (SRNS), Alport syndrome and rare inherited renal disease.

24 Sen ES, Dean P, Yarram-Smith L, *et al.* Clinical genetic testing using a custom-designed steroid-resistant nephrotic syndrome gene panel: analysis and recommendations. *J Med Genet* 2017;**54**:795–804. doi:10.1136/jmedgenet-2017-104811

25 Bierzynska A, McCarthy HJ, Soderquest K, *et al.* Genomic and clinical profiling of a national nephrotic syndrome cohort advocates a precision medicine approach to disease management. *Kidney Int* 2017;**91**:937–47. doi:10.1016/j.kint.2016.10.013

26 Feero WG, Guttmacher AE. Genomics, personalized medicine, and pediatrics. *Acad Pediatr* 2013;**14**:14–22. doi:10.1016/j.acap.2013.06.008

27 Hawcutt DB, Francis B, Carr DF, *et al.* Susceptibility to corticosteroid-induced adrenal suppression: a genome-wide association study. *Lancet Respir Med* 2018;**0**. doi:10.1016/S2213-2600(18)30058-4

28 Yu H, Burton B, Huang C-J, *et al.* Ivacaftor potentiation of multiple CFTR channels with gating mutations. *J Cyst Fibros* 2012;**11**:237–45. doi:10.1016/j.jcf.2011.12.005

29 Wainwright CE, Elborn JS, Ramsey BW, *et al.* Lumacaftor–Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del *CFTR*. *N Engl J Med* 2015;**373**:220–31. doi:10.1056/NEJMoa1409547

30 Lowe R, Shirley N, Bleackley M, *et al.* Transcriptomics technologies. *PLoS Comput Biol* 2017;**13**:1–23. doi:10.1371/journal.pcbi.1005457

31 Park J, Shrestha R, Qiu C, *et al.* Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* 2018;**360**:758–63. doi:10.1126/science.aar2131

32 Vickovic S, Magnusson J, Giacomello S, *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. 2016;**353**.

33 Mckinney EF, Lee JC, Jayne DRW, *et al.* T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection. Published Online First: 2015. doi:10.1038/nature14468

34 Banchereau R, Hong S, Cantarel B, *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* 2016;**165**:551–65. doi:10.1016/j.cell.2016.03.008

35 Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature* 2016;**537**:347–55. doi:10.1038/nature19949

36 Geyer PE, Kulak NA, Pichler G, *et al.* Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst* 2016;**2**:185–95. doi:10.1016/j.cels.2016.02.015

37 Keshishian H, Burgess MW, Gillette MA, *et al.* Multiplexed, Quantitative Workflow for Sensitive Biomarker Discovery in Plasma Yields Novel Candidates for Early Myocardial Injury. *Mol Cell Proteomics* 2015;**14**:2375–93. doi:10.1074/mcp.M114.046813

38 Anink J, Van Suijlekom-Smit LWA, Otten MH, *et al.* MRP8/14 serum levels as a predictor of response to starting and stopping anti-TNF treatment in juvenile idiopathic arthritis. *Arthritis Res Ther* 2015;**17**:200. doi:10.1186/s13075-015-0723-1

39 Gerss J, Roth J, Holzinger D, *et al.* Phagocyte-specific S100 proteins and high-sensitivity C reactive protein as biomarkers for a risk-adapted treatment to maintain remission in juvenile idiopathic arthritis: a

comparative study. *Ann Rheum Dis* 2012;**71**:1991–7. doi:10.1136/annrheumdis-2012-201329

40 Foell D, Wulffraat N, Wedderburn LR, *et al.* Methotrexate Withdrawal at 6 vs 12 Months in Juvenile Idiopathic Arthritis in Remission<subtitle>A Randomized Clinical Trial</subtitle> *JAMA* 2010;**303**:1266. doi:10.1001/jama.2010.375

41 Beck LH, Bonegio RGB, Lambeau G, *et al.* M-Type Phospholipase A $_2$ Receptor as Target Antigen in Idiopathic Membranous Nephropathy. *N Engl J Med* 2009;**361**:11–21. doi:10.1056/NEJMoa0810457

42 Hofstra JM, Wetzels JFM. Phospholipase A2 receptor antibodies in membranous nephropathy: unresolved issues. *J Am Soc Nephrol* 2014;**25**:1137–9. doi:10.1681/ASN.2014010091

43 Wishart DS, Jewison T, Guo AC, *et al.* HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res* 2012;**41**:D801–7. doi:10.1093/nar/gks1065

44 Clish CB. Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harb Mol case Stud* 2015;**1**:a000588. doi:10.1101/mcs.a000588

45 Fanos V, Cristina Pintus M, Lussu M, *et al.* Urinary metabolomics of bronchopulmonary dysplasia (BPD): preliminary data at birth suggest it is a congenital disease. *J Matern Neonatal Med* 2014;**27**:39–45. doi:10.3109/14767058.2014.955966

46 Horgan RP, Broadhurst DI, Walsh SK, *et al.* Metabolic Profiling Uncovers a Phenotypic Signature of Small for Gestational Age in Early Pregnancy. *J Proteome Res* 2011;**10**:3660–73. doi:10.1021/pr2002897

47 McLaren W, Gil L, Hunt SE, *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 2016;**17**:122. doi:10.1186/s13059-016-0974-4

48 ENCODE: Encyclopedia of DNA Elements – ENCODE. https://www.encodeproject.org/ (accessed 21 May 2018).

49 Rogers S, Girolami M, Campbell C, *et al.* The Latent Process Decomposition of cDNA Microarray Data Sets. *IEEE/ACM Trans Comput Biol Bioinforma* 2005;**2**:143–56. doi:10.1109/TCBB.2005.29

50 Luca B-A, Brewer DS, Edwards DR, *et al.* DESNT: A Poor Prognosis Category of Human Prostate Cancer. *Eur Urol Focus* 2017;**0**.

doi:10.1016/j.euf.2017.01.016

51 Research - Research - Medical Research Council. https://mrc.ukri.org/research/initiatives/stratified-medicine/research/ (accessed 20 May 2018).

52 Stubbs FE, Birnie MT, Biddie SC, *et al.* SKOV3 cells containing a truncated ARID1a protein have a restricted genome-wide response to glucocorticoids. *Mol Cell Endocrinol* 2018;**461**:226–35. doi:10.1016/j.mce.2017.09.018

53 Richards S, Aziz N, Bale S, *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;**17**:405–23. doi:10.1038/gim.2015.30